Novel Calibrations for FT-NIR Analysis of Protein, Oil, Carbohydrates and Isoflavones in Foods

Jun Guo*, Tiefeng You* and I.C. Baianu*,**

**AFC- NMR and NIR Microspectroscopy Facility, FSHN Dept., College of ACES, University of Illinois at Urbana, 305 Burnsides Research Laboratory, 1208 W. Pennsylvania, Ave., Urbana, IL. 61801*

*** NPRE Department, University of Illinois at Urbana, Urbana, IL.61801*

1. Introduction

The determination of food composition is fundamental to theoretical and applied investigations in food science and technology, and is often the basis of establishing the nutritional value and overall acceptance from the consumer standpoint. Most of the methods described in a previous report [1] (Baianu et al, 2011a) are useful for the conventional analysis of foods, that is, the determination of the major components (proteins, lipids, moisture, carbohydrates, and minerals). These components are included in standard tables of food composition. Advances in food analysis in the last three decades have resulted from the development of many instrumental methods such as NIR and from the improvements in separation methods (mainly chromatography).

The analyst often assumes that the sample to be analyzed is homogeneous. It is advisable that before starting a determination, the whole sample be mixed to eliminate heterogeneity – mainly in particle size and moisture distribution (Pomeranz and Meloan 1994). In some foods like concentrated sugar solutions, the sample must be heated carefully to dissolve sugar crystals.

Why would one wish to analyze the composition of soy and other health foods?

Because soy and other health foods are important for lowering cholesterol and the prevention, or treatment of atherosclerosis and coronary heart disease. Soy food composition is also important for weight loss/weight control (Liu et al., 1995). Therefore, quality control and routine monitoring of soy and other health food composition is important to the consumers. Monitoring the levels of isoflavones in health foods such as soymilk appears also to be important in populations that are at risk for certain types of cancers. Rapid, accurate, and cost-effective composition analyses of soyfoods and other health foods are essential for improving the efficiency and quality of health food production. This is the first attempt at developing Fourier

Transform Near Infrared Reflectance Spectroscopy (FT-NIRS) calibrations for soy and other health foods.

Soy tofu is a traditional soyfood originated from China (Liu et al. 1995). During the course of soybean cultivation, the Chinese had gradually transformed soybeans into various forms of soyfoods, including tofu, soymilk, soy paste, soy sauce and soy sprouts. Along with soybean cultivation, methods of soyfood preparation were gradually spread to Far East and West countries. The art of preparing soyfoods has now spread to the rest of the world, due to agricultural innovation and cultural exchanges. For the past several decades, advances in soybean chemistry and innovation in processing and packaging technology have dramatically modernized traditional ways of preparing soyfoods. As new medical research unveils the health benefits of soyfoods, such as the benefits of isoflavones for women's health, there is no doubt that soyfoods will soon become a part of global culture.

It is well known that protein is the dominant component in tofu. In an early report (Koga et al., 1992), reported that the NIR spectrum of tofu from 1100 to 2500 nm region was correlated with moisture, crude protein, and fiber contents determined by standard chemical methods, with correlation coefficients of 0.976, 0.830, and 0.865, respectively. Some other researchers studied the contributions of the total soybean proteins, the storage proteins [glycinin (11S) and b-conglycinin (7S) fractions] to tofu yield and texture. They analyzed protein contents by using SDS-PAGE (SDS-PAGE) coupled with densitometry and reversed phase-high performance liquid chromatography (RP-HPLC) (Mujoo et al., 2003). In order to measure directly, rapidly and accurately, the soy protein in gels a special tofu calibration was developed with a Spectrum One NTS FT-NIRS instrument.

Soymilk is another popular liquid soyfood, in which protein, carbohydrates and water are the three main components (Liu et al. 1995). Protein content in soymilk is usually determined by conventional methods such as chemical analysis and UV-Vis Spectroscopy method (Nielsen 1994). In a previous research on capillary electrophoresis, quantitation of bovine whey proteins in commercial powdered soybean milk was performed by adding bovine whey to its formulation using the calibration method of the external standard (Garcia-Ruiz et al., 1999). These techniques are either time-consuming, or not accurate enough for practical applications. A novel calibration was thus developed here with the Spectrum One NTS FT-NIR instrument to accurately measure protein, fat and carbohydrate contents in soymilk. For such a purpose, a transflectance working mode was employed for spectral data acquisition of soymilk. This mode is usually used for thin layer samples in order to reduce the noise level and baseline shift of spectra. If the NIR spectra

of liquid samples such as milk are obtained with the regular transmittance or reflectance mode, accurate quantitation is almost impossible because of the low S/N ratio caused by light scattering and large baseline shift (Ozaki et al., 2001).

The high dietary intake of soya has been associated with a reduced risk of some cancers such as breast cancer for women and heart disease. Isoflavones (mainly including daidzein, genistein and genistin) may be responsible for the protective role of soya (Liu 1997; Song et al. 1998). Monitoring the levels of isoflavones in health foods such as soymilk appears also to be important in populations that are at risk for certain types of cancers (Liu et al. 1995). Rapid, accurate, and cost-effective composition analyses of soy isoflavones are essential for breeding and genetic selection studies aimed at optimizing soybean seed compositions for human health food applications (Choi et al. 2000; Lee et al. 2003), and improving the efficiency and quality of soy health food production. The determination of isoflavones content is commonly done by HPLC analysis (Carrao et al. 2002; Choi et al. 2000; de-Rijke et al. 2001; Lee et al. 2003; Song et al. 1998; Tekel et al. 1999), or other improved methods with regular liquid chromatography (Kao et al. 2002). The HPLC method for isoflavone measurement is expensive, time-consuming, and impractical for measurements of large number of soybean samples that are required by breeding and selection studies.

Few studies on NIR, however, have been reported on the analysis of one or two seeds and no work has been published on measurement of low-level components such as isoflavones, mainly because of the limited spectral resolution and stability of conventional NIR instruments. In the past five years, significant improvements in NIR instrumentation have been achieved through applications of novel technologies such as Diode Array and Fourier Transform (Guo et al., 2002); which thereby provided the potential for single seed analysis of both major components and low-level components of soybeans. In this chapter, rapid and accurate analytical methods for protein, oil, moisture, and isoflavone determinations were developed with state-of-the-art FT-NIR instruments. This is the first attempt at developing Fourier Transform Near Infrared Reflectance Spectroscopy (FT-NIRS) calibrations for isoflavones in soybeans.

2. Calibration and validation methods

In this section partial least-squares regression models are employed to develop FT-NIR calibrations for soy and other health foods, soy tofu and milk, as well as soy isoflavones.

The general procedures for calibration development for the Spectrum One NTS can be described as:

1. Data acquisition with standard calibration samples.
2. Wavelength range selection suitable for sample composition determination, based on effective absorption bands.
3. Use “Interactive Baseline Correction” function to correct spectrum baselines and then normalize the spectra.
4. Matrix calculations by PLS-1 algorithm in order to optimize the calibration parameters after corrections for light scattering effects in raw spectra by MSC method.
5. Generate a calibration file with the optimized calibration parameters and make it ready for sample measurement.

In order to improve accuracy and robustness for calibration development, spectral data sets based on equally wide concentration ranges of all components were taken to statistically maximize the information content in the spectra (Haaland and Thomas, 1988). Despite that the calibration algorithm was designed for PLS-1 simulations with a three component mixture system, it is applicable to real samples with multiple components. Thus, wide concentration ranges of all components in standard samples are necessary for high quality calibration development. Although the concentration ranges of all components for real systems may not be comparable, it is necessary to make the concentration range of each component as wide as possible.

2.1. Calibration algorithm

2.1.1. Determining the Number of Factors for the Model

PLS-1 is a reduced subset of the full PLS-2 algorithm. The algorithms have been combined here, with appropriate notes on what they differ. Note that a PLS-2 model of a training set with only one constituent is identical to a PLS-1 model for the same data. One of the more subtle tasks in using PCR and PLS is choosing the correct number of loading vectors (factors) to use to model the data. As more and more vectors are calculated, they are ordered by the degree of importance to the model (either by variance in PCA or concentration weighted variance in PLS). Eventually the loading vectors will begin to model the system noise (which usually provides the smallest contribution to the data).

The earlier vectors in the model are most likely to be the ones related to the constituents of interest, while later vectors generally have less information that is useful for predicting

concentration. In fact, if these vectors are included in the model, the predictions can actually be worse than if they were ignored altogether. Thus, decomposing spectra with these techniques and selecting the correct number of loading vectors is a very effective way of filtering out noise.

However, if too few vectors are used to construct the model, the prediction accuracy for unknown samples will suffer since not enough terms are being used to model all the spectral variations that compose the constituents of interest. Therefore, it is very important to define a model that contains enough vectors to properly model the components of interest without adding too much contribution from the noise.

Models that include noise vectors or more vectors than are actually necessary to predict the constituents'concentrations are called overfit. Models that do not have enough factors in them are known as underfit.

Unfortunately, there is usually no clear indicator of how many factors are required to move from "constituent" vectors into "noise" vectors and prevent both underfitting and overfitting. However, there are a variety of methods that can be used to aid in determining this value. One of the most effective is to calculate the PRESS (Prediction Residual Error Sum of Squares) for every possible factor. This is calculated by building a calibration model with a number of factors, then predicting some samples of known concentration (usually the training set data itself) against the model. The sum of the squared difference between the predicted and known concentrations give the PRESS value for that model.

$$PRESS = \sum_{i=1}^{n} \sum_{j=1}^{m} (CP_{i,j} - C_{i,j})^2 \qquad \text{Eq. (2.1)}$$

In the above equation, n is the number of samples in the training set, and m is the number of constituents. Cp is the matrix of predicted sample concentration from the model, and C is the matrix of known concentrations of the samples. The smaller the PRESS value, the better the model is able to predict the concentration of the calibrated constituents. By calculating the PRESS value for a model using possible factors and plotting the results, a very clear trend should emerge.

2.1.2. Cross validation

Cross validation is conceptually very simple to understand, but it is also the most calculationally intensive method of optimizing a model. In effect, cross-validation attempts to emulate predicting "unknown" samples by using the training set data itself. The procedure is as follows:

1. Select a sample (or a small group of samples, if the training set is large enough) and remove the spectrum (spectra) and corresponding concentration data from the data matrix. Set the factor counter to I =1.
2. Use the remaining spectra and concentration data of the samples to perform the decomposition and calibration calculations for factor I (loading factor).
3. Predict the concentrations of the removed samples(s) using the calibration equation from step 2, and calculate PRESS(I).
4. Increase the factor counter (I=I+1) and repeat from step 2 until all desired factors (I=f) have been calculated and predicted.
5. Place the previously left our sample data back into the training set and select a different sample (or group). Return to step 1 and repeat the calculations. As each sample is left out, add the calculated squared residual error to all the previous PRESS values. Repeat until all samples have been left out and predicted at least once.

There are two main advantages of cross-validation over all other methods. The first is in how it estimates the performance of the model. Since the predicted samples are not the same as the samples used to build the model, the calculated PRESS value is a very good indication of the error in the accuracy of the model when used to predict "unknown" samples in the future. The larger the training set and the smaller the groups of samples left out in each pass (optimally only one sample at a time, but this can be very time consuming), the better this estimate will be. In effect, the model is validated with a large number of "unknown" samples (since each training sample is left out at least once) without having to measure an entirely new set of data.

The second benefit of cross validation is better outlier detection. While this will be discussed in more depth in a later section, it can be mentioned that cross validation is the only validation method that can give complete outlier detection for the training set data. Since each sample is left out of the models during the cross validation process, it is possible to calculate how well the spectrum matches the model by calculating the spectral reconstruction and comparing it to the original training spectrum (via the spectral residual). If the predicted concentrations for a

single sample are way off and the spectrum does not match the model very well but the rest of the data works very well, the sample is possibly an outlier. Identifying and removing outlier samples from the training set should always improve the predictive ability of the model. Only if a complete cross validation is performed, the outlier detection on the training set data can be well performed. Unfortunately, cross validation is a very time consuming process. It requires recalculating the models for every sample left out. However, there are a few somewhat acceptable shortcuts. If the number of samples in the training set is large enough, the number of samples rotated out in each pass can be more than one. This obviously does not give the best statistics for each sample, but it does speed the calculations and can be acceptable for determining the number of factors for the model.

2.1.3. Selecting the Factors Based on SECV

To avoid building a model that is either overfit or underfit, the number of factors where the PRESS plot reaches a minimum would be the obvious choice of the best model (except in the case of Self-Prediction). While the minimum of the PRESS may be the best choice for predicting the particular set of samples, it is not always optimum for prediction of all unknown samples in the future.

The concepts of *SECV* (*Standard Error of Cross Validation*) or *SEP* (*Standard Error of Prediction*) can be better uyilizedd to select the optimal number of factors, instead of PRESS. The definition of SECV is:

$$\mathbf{SECV} = \sqrt{\frac{\sum_{i=1}^{n} (Y_{i(k)} - Y_{i(p)})^2}{n}} \quad ,$$

Eq. (2.2)

where $Y_{i(k)}$ is the known concentration, $Y_{i(p)}$ is the predicted concentration and **n** is the number of samples calculated.

One notes that the SECV expression in Eq. (2.2) is comparable to PRESS. SECV is the averaged root mean square of PRESS, and thus it follows the same tendency of variation as PRESS does (in ThermoNicolet's TQ Analyst program, SECV is also called RMSECV, with RM standing for the root mean). When PRESS reaches its minimum, SECV reaches its minimum as well. However, the SECV represents the prediction error for building the calibration model better than PRESS does. Therefore, one may use SECV plots and values to indicate the optimized number

of factors as the choice for the best model. However, for a calibration that is required to be both robust and accurate, it is customary to choose the number of factors corresponding to the minimum in the plot of Log (PRESS) against the number of factors.

In Figure 2.1, which is the SECV vs. factor plot for soy tofu calibration development, one notices that for numbers of factors between 0 and 15 the SECV decreases as each new factor is added to the model. This indicates that the model is underfit and there are not enough factors to completely account for the constituents of interest as long as the SECV decreases significantly. At some point, the SECV plot should reach a minimum (6), and then begin to increase again. At this point the model is beginning to add factors that contain uncorrelated noise which are not related to the constituents of interest. When these extra "noise" vectors are included in the model, it is an overfit and its predictive ability is diminished. The number of factors for the minimum SECV value, (e.g. n = 6), is thus the best choice for the prediction. The correlation for calculated (predicted) protein percentage *vs*. actual protein percentage with 6 factors was plotted in Figure 2.2, and a correlation coefficient of 0.999 was reached.

2.1.4. Outlier sample detection

Outlier detection is equally important as choosing the optimum number of factors for the model. If one or more of the training samples are in error, it will cause errors in the calibration model and ultimately poor prediction results for unknowns. Outlier samples usually arise from some incorrect measurement, whether it is in the concentration data (i.e. errors in the primary calibration techniques, transcription errors), or in the spectral data (i.e. spectrometer error, sample handing procedures, environmental control such as temperature, humidity, etc.). Including outlier samples in the training set will introduce a bias to the final model. In effect, outlier samples will tend to "pull" the model in their direction, causing the predicted concentrations of valid samples to be less accurate (or even erroneous) than if the sample was completely eliminated from the training set.

Samples that have significantly larger concentration residuals (difference between the actual and predicted concentrations) than the rest of the training set are known as concentration outliers. This type of outlier generally arises when the experimenter either makes a mistake in creating the calibration mixtures or there was an error in the analysis of the samples from the primary calibration techniques used to generate the calibration concentration values. Another possibility which frequently occurs is a transcription error: the analyst simply types in the wrong

concentration value when building the computerized training set. Some obvious outliers can be simply picked up by visual inspection. While the human eye is excellent at discerning patterns in data, visual inspection is not always a valid basis for a decision of this type. What is really needed is a mathematical way to accurately determine the likelihood that a sample is really an outlier. For clusters of data points, it is possible to use a measure of the Mahalanobis distance (Mahalanobis, 1936). This is calculated as the distance of the potential outlier sample point as measured from the mean of all the remaining points in the cluster. The distance is scaled for the range of variation in the cluster in all dimensions, and then assigns a probability weight to the sample in terms of standard deviation. Any sample which lies outside of 3 standard deviations from the mean can be considered suspicious, e.g. 3% deviation for soy and health food composition. The Mahalanobis distance is also useful in qualitative analysis of spectral data for which the constituent concentrations are not known.

2.2. Spectra pre-processing

One of the major problems in applying chemometric models to spectra is the fact that the acquired spectrum of a sample is dependent on many different, sometimes uncontrollable factors. For example, samples of powdered solids are usually measured by diffuse reflectance. Light scattering off the particles causes every spectrum, even remeasurement of the same sample, to

Figure 2.1. SECV vs. factor plot for protein in the calibration development for soy tofu

Figure 2.2. Calculated (or predicted) protein% vs. Actual (or reference) protein% plot, with 6 factors, in the calibration development for soy tofu.

look a little bit different due to the particle size distribution and alignment with the incident beam of light. While the quantitative information related to the constituents is still contained within the spectral data, it may not be immediately apparent. Another example is that the pathlength of the samples sometimes can not be controlled, such as measuring spectra of thin films.

Chemometric models can sometimes correct for these effects by adding extra loading vectors, but generally the models will perform better if they can be removed or at least minimized before running the data through the calculations. Since they are applied to the data before it is used in the model, they are often called Preprocessing Algorithms. There are a variety of methods that can be used to remove the non-constituent related aberrations in the data. Most algorithms are targeted at removing a specific interference (*MSC*, for example, specifically attempts to remove the effects of light scattering). Properly applying preprocessing requires understanding the interference in the data and selecting the appropriate algorithms to correct the effects.

2.2.1 Multiplicative Scatter Correction (MSC)

The NIR detector receives light coming from the sample in form of: diffuse reflectance after absorption, specular reflection and scattered light. Only the diffuse reflectance contains chemical composition information, whereas the latter two do not. Therefore, in order to determine accurately chemical composition from NIR measurements, the light scattering and specular components must be corrected for (Williams and Norris, 1987).

The degree of scattering is dependent on the wavelength of the light that is used, and not uniform throughout the spectrum. Typically, this appears as a baseline shift, tilt and sometimes curvature. It is not simply a matter of measurement errors that light scattering effect may cause. In an early research about scatter-correction for NIR reflectance spectra of meat (Geladi et al., 1985), reflectance for fat shows completely different tendencies (up and down) before and after MSC correction (see Figure 9 on page 498 of Geladi et al., 1985). Therefore, without MSC correction, the raw reflectance or absorbance values will make a totally incorrect calibration, and lead to wrong prediction for unknown samples. The MSC method assumes that the wavelength dependency of the light scattering is different from that of the constituent absorption. Theoretically, by using data from many wavelengths in the spectrum, it should be possible to separate the two.

This method attempts to remove the effects of scattering by linearizing each spectrum to some "ideal" spectrum of the sample (Galactic, 1996). MSC begins with a calculation of the *average spectrum* from all the data in the training set and uses it as the "ideal" spectrum. Thereafter, the spectral responses in each spectrum are used to calculate a linear regression against the corresponding points in the ideal spectrum. The slope and offset values from this regression are subtracted and ratioed respectively in the original training spectrum to give the MSC corrected spectrum.

$$\overline{A_j} = \sum_{i=1}^{n} A_{i,j}$$

which is the *Mean Spectrum*: Eq. (2.3)

Linear Regression: Eq. (2.4)

$$A_i = m_i \overline{A} + b_i$$

MSC Correction:

$$A_{i(MSC)} = (A_i - b_i) / m_i$$

Eq. (2.5)

In these equations, ***A*** is the ***n*** by ***p*** matrix of training set spectral responses for all the wavelengths, **A-bar** is a **l** by ***p*** matrix of the average responses of all the training set spectra at each wavelength, $\mathbf{A_i}$ is a **l** by ***p*** matrix of the responses for a single spectrum in the training set, n is the number of training spectra, and p is the number of wavelengths in the spectra. The $\mathbf{m_i}$ and b_i values are the slope and offset coefficients of the linear regression of the mean spectrum vector **A-bar** versus the $\mathbf{A_j}$ spectrum vector. By adjusting the slope and offset of the sample spectra to the "ideal" average spectrum, the chemical information is preserved while the differences between the spectra are minimized. Thus, the major source of random variance between them can be removed as much as possible.

2.2.2. Correcting Baseline Effects:

Spectrometers rarely collect data with an ideal baseline. In order to accurately calculate concentrations, it is necessary to remove the baseline shift effect introduced by the spectrometer, especially by specular reflectance in the reflectance mode for PerkinElmer's Spectrum One NTS.

There are a number of methods used by spectroscopists to remove baseline effects from the spectra they collect. The problem with most methods is that they require the spectroscopist to decide that the baseline is corrected by visual inspection. However; there are some methods which are reasonably automated enough to be used as part of a calibration model, such as Linear Regression Baseline Fitting, Two Point Linear Baseline approach, and Derivatives. In PerkinElmer's Spectrum program, a special function "Interactive Baseline Correction" is designed for users to correct baseline shift for raw spectra, and another function "Normalization" is used to normalize spectra so that the absorbance values can be used correctly to fit Beer's Law for matrix calculations.

2.3. Computer iteration steps for calibration development with PLS-1

The calibration involves regression with a *Partial Least Squares* Type **1** (*PLS-1*), multi-variate algorithm (Galactic Industries Corporation, 1996). The collection of known data, or chemical composition, for each standard samples, together with the measured data by the instrument are called a calibration set (or training set). Such calibration algorithms as PLS-1 base their predictions of each constituent concentration on changes in the spectral data rather than absolute absorbance values. A simpler algorithm called "NIPALS" is useful to illustrate the iteration procedures followed in PLS-1 as well. The NIPALS algorithm involves two stages: an iterative stage that utilizes just the NIR spectral data and a regression stage that utilizes the laboratory composition data along with the results from the previous stage**.** The first iteration stage begins by computing the difference between each raw spectrum and the mean spectrum, $A_i - \overline{A}$, for the entire calibration set. A set of factors F, or eigenvectors F_i are then iterated by setting such factors at the beginning to be equal to the raw spectra, A_i. Both A and F are represented as tables (or matrices) of the NIR absorbance values at specific wavelengths across the NIR spectrum of soybeans. From these matrices, one calculates tables (or matrices) of scores, S_i, defined as a product of two matrices:

$$S_i = A_i F_i' \qquad \text{Eq. (2.6)}$$

where $\mathbf{F_i'}$ is the transposed matrix of the eigenvector F_i . In a second iteration step, the eigenvectors Fi are normalized by dividing through the corresponding eigenvalues, $\lambda_{i,i}$, defined as:

$$\lambda_{i,i} = (\Sigma S_i^2)^{1/2} \qquad \text{Eq. (2.7)}$$

Thus $\mathbf{F_i} = \mathbf{A_i} / \lambda_{i,i}$ are the normalized eigenvectors at this second iteration step. A new set of scores is then calculated with equation (2.6) from the normalized eigenvectors. The new set of scores is subtracted from the corresponding ones obtained at the first iteration step. The iteration is complete when this difference is zero or negligible. If the difference is significant, one re-iterates the eigenvectors F_i through matrix multiplication:

$$F_i = (Ai - \overline{A})' \times S_i \quad , \qquad \text{Eq. (2.8)}$$

until the difference between two values of S_i for consecutive iterations becomes zero or negligible. Such optimized scores are effectively the absorbance values of individual constituents at selected wavelengths across the NIR spectrum of the soybeans.

The tables of those score values obtained at the first stage are then employed in a second stage to relate the absorbance values of individual constituents to the known chemical composition stored as a chemical composition table, or matrix, **C**. The model equation at this stage is therefore:

$$C = B \times S + E_c \qquad \text{Eq. (2.9)}$$

where **B** is regression coefficient matrix and $\mathbf{E_c}$ is a matrix table of regression error terms for chemical composition of the constituents. Once the regression coefficients in martix **B** are determined, the calibration is complete and can be utilized to predict composition values for the constituents of unknown samples.

In the PLS-1 algorithm, an added sophistication is introduced by utilizing from the first pass of the iteration a linear combination of calibration spectra weighted by the corresponding concentrations of one constituent at a time. In this procedure, the loading vectors (sometimes called “spectral weighing vectors”), are defined as:

$$\mathbf{W_j = C_j' A} \qquad \text{Eq. (2.10)}$$

where $\mathbf{C_j}$ is the composition vector for constituent ***j***. At the next iteration pass, these spectral weighing vectors are normalized as follows:

$$\mathbf{W_j\ (pass\ 2) = W_j\ (pass\ 1) / [\ W_j\ (pass1)\ W_j'\ (pass\ 1)]} \qquad \text{Eq. (2.11)}$$

Therefore, by using loading vectors as eigenfactors, concentration information is included in the calculations during the first spectral decomposition stage rather than in a separate second stage. This is the main difference between PLS and the NIPALS (also the PCR method).

Loading factors are actually mimics of the pure component spectra. The first loading factor in the PLS-1 analysis is a first-order approximation to the pure-component spectrum of the corresponding component. Figure 2.3 gives one graph of the first loading factors for the pure components in SPI and H_2O mixture. The pure component spectra of SPI and H_2O generated by the computer program look exactly the same as their real spectra.

The number of calibration loading factors for each constituent can be obtained for the minimum value of the SECV. However, for a calibration that is required to be both robust and accurate, it is customary to choose the number of factors corresponding to the minimum in the plot of Log (PRESS) against the number of factors.

2.4. Standard Error of Prediction (SEP)

Standard Error of Prediction (SEP) has the same definition as SECV, but the samples for SEP are not involved in the cross validation process for calibration development. The samples for SEP are only used to compare predicted values from the developed calibration with known values for calibration validation purposes.

3. Experimental results and data analysis

3.1. NIR analysis of soy and other health foods

3.1.1. Sampling and experiments

FT-NIRS measurements were carried out in quadruplicate for 16 types of food samples,such as: soy crisps, dry roasted soy nuts, soy burgers, soy tofu, island black beans, soymilk powder, rye cakes, rye bread, rye toast, rye cocktail bread, dry tomato, popcorn minicakes, biscuits and lean ham.

Figure 2.3. ***A graph of loading factors for the pure components in SPI and H_2O mixture***.

Their composition values were calculated according to the nutrition tables on those products and used for calibration data, which are listed in **Table 1.** The other standard samples were prepared by either dehydrating or rehydrating some of the original samples. The total number of samples used for this calibration development was 28. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with a PerkinElmer Co.'s FT-NIR spectrometer, model Spectrum One NTS NTS. This spectrometer is optimized for high-sensitivity analysis of solid samples, being equipped with an NIRA, integrating sphere accessory and an extended range InGaAs detector. The beam size was set to be 8.94 mm. The number of scans was 64 for each spectrum.

Table 1. *Composition values of 16 soy and other health foods calculated according to the nutrition tables on those products*.

	Protein%	**Fat%**	**Moisture%**	**Total Carbohydrates%**	**Fiber%**
Soy crisps	25.0	7.1	0.5	50.0	7.1
Dry roasted soy nuts	43.3	26.7	< 1.0	20.0	13.3
Soy burgers	20.0	4.4	59.7	8.9	5.6
Frida's firm tofu	7.1	3.5	82.2	2.4	1.2
Fried tofu	9.1	8.6	76.1	2.0	1.0
Island black beans	18.8	1.6	5.0	53.1	18.8
Soymilk powder	10.0	5.9	1.7	69.1	0.1
Popcorn minicakes	12.5	6.3	< 1.0	75.0	6.3
Rye cakes	13.0	< 0.1	1.0	60.0	26.0
Rye bread	9.7	4.8	26.0	45.2	6.5
Light rye bread	7.3	3.7	22.0	48.8	2.4
Rye toast	10.0	< 0.1	< 1.0	85.0	3.0
Biscuits	10.0	< 0.1	< 1.0	85.0	2.0
Dry tomato	<1.0	< 0.1	4.0	86.0	9.0
Rye cocktail bread	9.7	4.8	26.0	45.2	6.5
Bohllen lean ham	14.5	14.2	68.8	1.8	< 0.1

3.1.2. Calibration results

The TQ Analyst software developed by Nicolet Instruments was employed to process NIR spectra and develop calibration files. A total of 112 FT-NIR spectra were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments. Figure 2.4. shows overlay of group spectra for soy and other health foods obtained with Spectrum One NTS after baseline correction and normalization. Standard composition values of major food components, such as: protein, fat, moisture, fiber, total carbohydrates were obtained from nutrition tables on those products. Composition changes of soy and other health foods caused by microwave heating or moisture rehydration were also monitored. The composition ranges for calibration development are: protein 0.5% to 43.3%, fat 0.1% to 26.7%, moisture 0.5 to 82.2%, fiber 0.1% to 26%, total carbohydrates 0.5% to 95%. These are quite wide concentration ranges and cover almost all of the soy and other health foods contents. The optimized parameters for the calibration result are listed below

Table 2. *Optimized SECV, R^2 values and number of factors for the calibrations developed on Spectrum One NTS, Wavelength range 4080 to 11200 cm^{-1}.*

	Protein%	Fat%	Moisture%	Total Carbohydrates%	Fiber%
SECV	1.2	0.7	1.4	1.7	1.0
R^2	0.992	0.994	0.995	0.995	0.985
# Factors	12	12	13	14	13
SEP	1.4	1.0	1.6	1.7	1.3

This calibration for soy and other health foods is characterized by low standard errors (~1%) and high degrees of correlation between NIR calculated values and laboratory reference values (~99%). It will satisfy commercial determination of nutritional contents in soy and other health foods. The purpose of developing this calibration is to introduce a new experimental method for rapidly and accurately measuring different types of soy and other health foods. The results were reported as (see Appendix) “Determination of Soy and Other Health Foods

Figure 4. *Overlay of FT-NIR Reflectance spectra for soy and other health foods obtained with Spectrum One NTS.*

(Source: Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, *Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean,* August 11-14, 2002, p506).

3.2. NIR analysis of soy tofu

3.2.1. Sampling and experiments

FT-NIRS measurements were carried out in quadruplicate for 19 tofu samples with different protein and water contents. The original tofu sample was a commercial product Fridas' Firm Tofu, with 7.1% protein, 82.2% water, and ~10% other total solid components such as fat, salts and carbohydrates. The other samples were prepared by short time microwave heating with an interval of 20 seconds, so that the water in tofu could be lost gradually and the protein content increased accordingly. The total number of samples used for this calibration development was 24. The composition values were calculated according to the amount of water loss. The composition ranges for calibration development are: protein 7.1% to 39.8%, moisture 27.1% to 82.2%, and other total solids 10.7% to 33.1%. These are quite wide concentration ranges and cover almost all of the soft and firm tofu contents. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set at 8.94 mm. The number of scans was 64 for each spectrum.

3.2.2. Calibration results

The TQ Analyst software was employed to process NIR spectra and develop calibration files. Totally 96 FT-NIR spectra (shown in Figure 2.5) were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments.

The optimized parameters for the calibration result are listed below:

Figure 5. *Overlay of FT-NIR Reflectance spectra for soy tofu obtained with Spectrum One.*

Table 3. *Optimized SECV, R^2 values and number of factors for the calibrations developed on Spectrum One NTS, wavelength range 4080 to 11000 cm^{-1}.*

	Protein%	**Moisture%**
SECV	0.75	1.19
R^2	0.999	0.998
# Factors	10	8
SEP	0.83	1.35

The calibration for soy tofu is characterized by low standard errors (~1%) and high degrees of correlation between NIR calculated values and laboratory reference values (~99%), and can be used to measure protein content in tofu.

3.3. NIR analysis of soy milk

3.3.1. Sampling and experiments

FT-NIRS measurements were carried out in quadruplicate for 27 soymilk samples with different protein and water contents. The liquid soymilk samples were made from a commercial soymilk powder product Mount Elephant Soybean Drink (Guangxi Cereal and Oil Product Company, Wuzhou City, Guangxi Province, China), with 10% protein and 69% carbohydrates. After mixing the soymilk powder with different portions of water, liquid soymilk samples were prepared for different concentrations. FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set to be 8.94 mm. The number of scans was 64 for each spectrum.

Owing to the fact that water is the dominant component in soymilk, protein bands on the soymilk spectra are overlapped by huge water bands. In order to get as much chemical information of the other components except water as possible, a specially designed metal reflector was used to obtain the transflectance spectra. Only 5 μl of liquid sample was put onto the instrument each time, with the reflector covered on top of the liquid layer, in order not to lose diffuse reflectance signals.

3.3.2. Calibration results

The TQ Analyst software was employed to process NIR spectra and develop calibration files. Totally 108 FT-NIR spectra (see **Figure 6**.) were preprocessed by applying a suitable

Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed to develop high-quality calibration models. The composition ranges for calibration development were: protein 0.5% to 10%, water 1.7% to 100%, and carbohydrates 3.5% to 69.1%. These are quite wide concentration ranges and cover almost all of the soymilk and even tofu contents. The optimized parameters for the calibration result are listed below.

(Figure 2.6.)

Table 4. Optimized SECV, R^2 values and number of factors for the calibration of soymilk developed on Spectrum One NTS, wavelength range 4080 to 11500 cm^{-1}.

	Protein%	**Fat%**	**H_2O %**	**Carbohydrates%**
SECV	0.03	0.02	0.34	0.23
R^2	0.999	0.999	0.999	0.999
# Factors	9	9	9	9
SEP	0.08	0.04	0.73	0.53

The calibration for soy milk is characterized by low standard errors, especially for protein and fat (<0.1%), and high degrees of correlation between NIR calculated values and laboratory reference values (~99%). It is suitable for measuring soymilk within regular concentration ranges. The results were reported as (see Appendix A) "Rapid Determinations of Soybean Isoflavones, Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, Proceedings of the China and International Soy Conference and Exhibition 2002 (CISCE 2002), November 6-9, 2002, 391-392.

Figure 6 (2.6). *Overlay of 108 FT-NIR Transflectance spectra for soymilk obtained with Spectrum One NTS*

3.4. NIR analysis of soybean isoflavones

3.4.1. Sampling and experiments

In order to develop NIR calibrations on such instruments for soybean composition analysis, soybean standard samples were selected from the USDA Soybean Germplasm Collection (Urbana, IL, USA). The selection of standard samples was based on their protein, oil, moisture, and isoflavone contents, to ensure that the ranges of standard sample constituent contents covered the full range of possible constituent variations of samples. To minimize screening effects of the soybean seed coat (especially black and brown coat) on isoflavones, soybean seeds were ground for preparation of standard samples.

Twenty eight ground soybean samples from isoflavone standards plus one isoflavone tablet sample (NovaSoy tablet) were utilized in the calibration development, with isoflavones range from 0.04% to 0.9% (HPLC data), protein range from 34% to 47.1% (ZX-50 data), oil range from 12.8% to 23% (ZX-50 data), and moisture range from 5.6% to 11.0% (ZX-50 data). Laboratory reference values of isoflavone composition were obtained by HPLC analyses of soybeans, which were kindly provided by Dr. J. Widholm's laboratory at UIUC. The TQ Analyst software was employed to process NIR spectra and develop calibration files. Totally 116 FT-NIR spectra were preprocessed by applying a suitable Multiplicative Scattering Correction (MSC). Partial Least Squares Type 1 (PLS-1) multivariate regression analyses were employed for high-quality calibration model developments.

The samples were ground with a Braun KSM2B Grinder. The average grinding time is 25 seconds, producing a powder sample with a particle size ranging from 100 um to 200 um. Quadruplet FT-NIRS measurements were carried out for the 29 isoflavone samples with a weight of 300 mg each (two soybean seeds). FT-NIR spectra were collected over a spectral range from 4000 to 12000 cm^{-1} (833 to 2500 nm) at a resolution of 8 cm^{-1} with Spectrum One NTS. The beam size was set at 8.94 mm. The number of scans was 32 for each spectrum.

3.4.2. Calibration results

Figure 7 (2.7) shows an overlay of FT-NIR spectra of ground soybeans for isoflavones standards. They are baseline corrected and normalized. A calibration was developed based on these spectra. Standard composition values were obtained with ZX-50 instrument for protein, oil, moisture and HPLC data for isoflavones.

The optimized parameters for the calibration result are listed below in Table 2.5. Figure 2.8 presents the calibration plot for calculated isoflavones% vs. actual (or reference) isoflavones %, with 9 factors. The correlation coefficient R and SECV (RMSEC) values are also listed in the figure.

Table 5. *Optimized SECV, R^2 values and number of factors for the calibration of soybean isoflavones developed on Spectrum One NTS, wavelength range 4100 to 10625 cm^{-1}.*

	Protein%	**Oil%**	**H_2O %**	**Isoflavones %**
SECV	0.67	0.28	0.12	0.0146
R^2	0.989	0.994	0.995	0.997
# Factors	6	9	9	9
SEP	0.43	0.32	0.26	0.0172

This calibration for soybean isoflavones is characterized by low standard errors (<0.02%) and high degrees of correlation between NIR calculated values and laboratory reference values (~99%). For soybean samples containing a normal isoflavone content, i.e. 0.2% to 0.9%, the calibration is accurately applicable. For soybean samples containing a low isoflavone content, i.e. 0.04% to 0.2%, the calibration can roughly predict the isoflavone concentration. The accuracy of this calibration is comparable with that of a recently published calibration for soybean isoflavones developed with single half soybean seeds (You et al., 2002). The results were reported as (see Appendix) (1) "Rapid Determinations of Soybean Isoflavones, Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy", Jun Guo and Ion C. Baianu, Proceedings of the China and International Soy Conference and Exhibition 2002 (CISCE 2002), November 6-9, 2002, 391-392.

Figure 7 (2.7.) *Overlay of FT-NIR Reflectance spectra for soybean isoflavone standards obtained with Spectrum One NTS*

Figure 8 (2.8.) *The calibration plot for calculated isoflavones% vs. actual (or reference) isoflavones%, with 9 factors.*

4. Conclusions

PerkinElmer's SpectrumOne NTS FT-NIR instrument can be utilized for accurate measurements of food protein, oil (fat), carbohydrate and fiber contents for both solid and liquid samples, as well as isoflavones contents in soybean powder samples. It can be also employed to obtain detailed characterization of foods and to investigate the interactions between major food components such as: protein, oil, water and carbohydrates. Moreover, fast and economical measurements of food composition allow online quality control and chemical analysis in food production.

REFERENCES

AACC. 1995. Approved Methods of Analysis, 9th ed. The American Association of Cereal Chemists. St. Paul, Minnesota.

AOAC International. 1995. Official Methods of Analysis, 16th ed., AOAC International, Gaithersburg, Maryland.

AOCS. 1998. Official Methods and Recommended Practices of the AOCS - 5th ed. The American Oil Chemists' Society. Champaign, Illinois.

Avram, M. and Mateescu, G.D. 1966. In *Infrared Spectroscopy Applications in Organic Chemistry,* 41-47.

Bach, K.E. and Li, B.W. 1991. Determination of oligosaccharides in protein-rich feedstuffs by gas-liquid chromatography and high-performance liquid chromatography. J. Agric. Food Chem. 39(4), 689-694.

Baer, R.J., Frank, J.F. and Loewenstein, M. 1983. Compositional analysis of whey powders using near infrared reflectance spectroscopy. *J. Food Sc*i. 48, 959.

Baianu, I.C. 1993. *Value Added Soybean Summi*t, Washington, D.C.

Baianu, I.C. and Kumosinski, T.F. 1994. Physical Chemistry of Food Processes. Vol. 2, 338-420.

Baianu, I.C., Cho, S.I., Stroshine, R.L. and Kruntz, G.W. 1993. Nondestructive sugar content measurements of intact fruit using spin spin relaxation time (T2) measurements by pulsed 1H magnetic resonance. Trans. ASAE. 36, 1217-1221.

Baianu I.C., Yakubu, P.I., Iozu, E.M. and Orr, P.H. 1993. Hydration of potato starch in aqueous suspensions determined from nuclear magnetic relaxation studies by oxygen, deuterium and proton NMR: Relaxation mechanisms and quantitative analysis. *J. Agric. Food Chem*. 41(2), 162-167.

Baianu I.C., Yakubu P.I. and Orr, P.H. 1990. Unique hydration behavior of potato starch as determined by deuterium NMR. *J. Food Sci.* 55(2), 458-461.

Baianu, I.C., You T., Guo, J. and Nelson, R.L. 2002. Calibration of Dual Diode Array and Fourier
Transform NIR Spectrometers for Composition Analysis of Single Soybean Seeds in Genetic
Selection, Cross Breeding Experiments, *Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean,* August 11-14, 2002, 508.

Baianu, I. C., You T., Costescu, D.M., Lozano, P.R., Prisecaru, V. and Nelson, R.L. 2004. Chapter 11, "High Resolution Nuclear Magnetic Resonance and Near Infrared Determination of Soybean Oil, Protein and Amino Acid Residues in Soybean Seeds", D. Luthria, ed., AOCS Publications. 193-248.

Babic, M., Turan, J., Babic, L. J. and Lazic, V. 2000. The influence of drying and storing technology onto soybean mechanical properties. PTEP (Yugoslavia). 4(3-4), 57-59.

Bartholomew, D.T., and Osuala, C.I. 1988. Use of the InfraAlyzer in proximate analysis of mutton. *J. Food Sci*. 53, 379.

Belitz, H.D. and Grosch, W. 1987. *Food Chemistry*, Springer-Verlag, Berlin.

Birth, G. S. 1986. The Light Scattering Characteristics of Ground Grains. Intern. Agrophys. 2(1), 59-67.

Birth, G.S., Dull, G.G., Renfroe, W.T. and Kays, S.J. 1985. Nondestructive spectrophotometric determination of dry matter in onions. J. Am. Horticultural Soc. 110, 297.

Brim, C.A. and Burton, J.W. 1979. Recurrent selection in soybeans (Glycine max): 2. Selection for increased percent protein in seeds. *Crop Sci*. 19(4), 494-498.

Buning, P.H. and Diller, M. 2000. Rapid analysis of foods using near-infrared spectrometry (NIRS): development, use and new possibilities. *Ernahrungs Umschau*. 47(1), 15-20.

Buslov, D.K. and Nikonenko, N.A. 1997. Regularized Method of Spectral Curve Deconvolution. *Appl. Spectrosc*. 51, 666.

Buslov, D.K., Nikonenko, N.A., Sushko, N.I. and Zhbankov, R.G. 1999A. Analysis of the results of α-D-glucose Fourier transform infrared spectrum deconvolution: comparison with experimental and theoretical data. *Spectrochimica Acta Part A*. 55, 229-238.

Buslov, D.K., Nikonenko, N.A., Sushko, N.I. and Zhbankov, R.G. 1999B. Deconvolution of Fourier transform infrared spectrum of α-D-galactose: comparison with experimental and theoretical data. *Spectrochimica Acta Part* A. 55, 1101-1108.

Cameron, A.G. 1967. An assessment of the potential application of the method of attenuated total reflectance (ATR) infrared qualitative analysis to food materials. *J. Food Technol.* 2, 223.

Carrao, P.M., de-Goes, F.S. and Kikuchi, A. 2002. Extraction time for soybean isoflavone determination. *Brazilian Archives Bio. Technol*. 45 (4), 515-518.

Cho, R.K., Iwamoto, M. and Saio, K. 1987. Determination of 7S and 11S Globulins in Ground Whole Soybeans by Near Infrared Reflectance Spectroscopy Analysis. *Nippon Shokuhin Kogyo Gakkaishi.* 34(10), 666-672.

Choi, Y.S., Lee, B.H., Kim, J.H. and Kim, N.S. 2000. Concentration of phytoestrogens in soybeans and soybean products in Korea. *J. Sci. Food Agric*. 80(12), 1709-1712.

Code of Federal Regulations. 1997. Title 21, Part 101- Food Labeling, *US Government Printing Office,* Washington DC.

Cox, R.J., Lebrasseur, E., Michiels, H. B., Li, H., Voort van-de, F.R., Ismail, A., Sedman, J. and Li, H. 2000. Determination of iodine value with a Fourier transform-near infrared based global calibration using disposable vials: an international collaborative study. *J. Am. Oil Chem. Soc.* 77(12), 1229-1234.

Cregan, P., Zhu, Y.L., Song, Q.J., Hofmann, N.E., Yaklich, R.W., Specht, J.E. and Nelson, R.L. 2002. Linkage disequilibrium and association analysis for QTL discovery. Program and Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean, August 11-14, 2002, P.106

de-Rijke, E., Zafra, G.A., Ariese, F., Brinkman, U.A. and Gooijer, C. 2001. Determination of isoflavone glucoside malonates in Trifolium pratense L. (red clover) extracts: Quantification and stability studies. *J. Chromatography A*. 932 (1-2), 55-64.

Delwiche, S.R. 1995. Single wheat kernel analysis by near-infrared transmittance: Protein content. *Cereal Chem*. 72, 11-16.

Dev, S.B., Rha, C.K. and Walder, F. 1984. Secondary structural changes in globular protein induced by a surfactant: Fourier self-deconvolution of FT-IR spectra. J. Biomolecular Structure & Dynamics. 2(2), 431.

Diem, M. 1993. *Introduction to Modern Vibrational Spectroscopy*, Chapter 6, John Wiley & Sons, Inc.

Diers, B.W., Keim, P., Fehr, W.R. and Shoemaker, R.C. 1992. RFLP Analysis of Soybean Seed Protein and Oil Content. *Theor. & Appl. Genetics*. 83, 608-612.

Dull, G.G., Birth, G.S., Smittle, D.A. and Lefler, R.G. 1989. Near infrared analysis of soluble solids in intact cantaloupe. *J. Food Sci*. 54, 393.

Espeja, E., Garcia, M., Marina, M. and Luisa, C. 2001. Fast detection of added soybean proteins in cow's, goat's, and ewe's milk by perfusion reversed-phase high-performance liquid chromatography. *J. Separation Sci*. 24(10/11), 856-864.

Fahrenfort, J. 1961. Attenuated total reflection. A new principle for the production of useful infrared reflection spectra of organic compounds. *Spectrochim. Acta* 17, 698.

Fox, J.D. and Robyt, J. F. 1990. Miniaturization of the three carbohydrates analyses using a microsample plate reader. Anal. Biochem. 195, 93-96.

Frankel, E.N., Nash A.M. and Snyder, J.M. 1987. A methodology study to evaluate quality of soybeans stored at different moisture levels. J. Am. Oil Chem. Soc. 64(7), 987-992.

Galactic Industries Corporation. 1996. GRAMS/32 User's Guide: PLS plus/IQ for GRAMS/32 and GRAMS/386, chapters 1-4.

Gandhi, A.P., Nenwani, M.M. and Ali, N. 1984. Investigations on the trypsin inhibitor, urease and cooking behavior of soybean Glycine max. Merr. Food Chem. 15(3), 215-218.

Garcia-Ruiz, C., Torre, M. and Marina, M.L. 1999. Analysis of bovine whey proteins in soybean dairy-like products by capillary electrophoresis. J. Chromatography A 859(1), 77-86.

Geater, C.W., Fehr, W.R. and Wilson, L.A. 2000. Association of Soybean Seed Traits with Physical Properties of Natto. Crop Sci. 40, 1529-1534.

Geater, C.W., Fehr, W.R., Wilson, L.A. and Robyt, J.F. 2001. A More Rapid Method of Total Sugar Analysis for Soybean Seed. Crop Sci. 41, 250-252.

Geladi, P., MacDougall, D. and Martens, H. 1985. Linearization and Scatter-Correction for Near -Infrared Reflectance Spectra of Meat. Appl. Spectrosc. 39(3), 491-500.

Greenspan, L., 1977. Humidity fixed points of binary saturated aqueous solutions. J. Res. of NBS. A. Phys. Chem., 81A(1), 89.

Guo J. and Baianu, I.C. 2002A. Determination of Soy and Other Health Foods Composition by Fourier Transform Near Infrared Reflectance Spectroscopy. Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean, August 11-14, 2002, P506.

Guo, J., You T., Baianu, I.C., Nelson, R.L. 2002B. Evaluation of Four Dispersive NIR Instruments and Methodology Development for Soybean Composition Analysis in Genetic Selection and Breeding Programs, Proceedings of the China and International Soy Conference and Exhibition 2002 (CISCE 2002), November 6-9, 2002, 412-413.

Haaland, D. M, and Thomas, E.V. 1988. Partial Least Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. Anal. Chem. 60, 1193-1202.

Harrick, N.J. 1960. Surface Chemistry from spectral analysis of totally internal reflected radiation. J. Phys. Chem. 64, 1110.

Haswell, S.J. 1992. Practical Guide to Chemometrics. Marcel Dekker, Inc., New York.

Hartwig E.E., Kuo, T.M. and Kenty, M.M. 1997. Seed Protein and Its Relationship to Soluble Sugars in Soybean, Crop Sci. 37, 770-773.

Hicks, K.B. 1988. HPLC of carbohydrates. Advances in Carbohydrate Chemistry and Biochemistry. 46, 17.

Himmelsbach, D. S., Barton II, F.E., McClung, A.M. and Champagne, E.T. 2001. Protein and apparent amylose contents of milled rice by NIR-FT/Raman spectroscopy. Cereal Chem. 78(4), 488-492.

Howard, M. 2001. Near-Infrared Applications in Biotechnology, (Biochemical and Pharmaceutical). Chapter 11 “Fundamentals of Near-Infrared Spectroscopy”, page 312, Edited by Raghavachari, R. Promega Corporation, Madison, Wisconsin.

Hymowitz, T., Collins, F.I., Panczner, J. and Walker, W.M. 1972. Relationship between the content of oil, protein, and sugar in soybean seed. Agron. J. 64, 613-616.

Hymowitz, T., Dudley, J.W., Collins, F.I. and Brown, C.M. 1974. Estimations of protein and oil concentration in corn, soybean, and oat seed by near-infrared light reflectance. Crop Sci. 14, 713-715.

Iowa Agriculture and Home Economics Experiment Station. 1990. Special Report 92.

Isaksson, T., Miller, C.E. and Naes, T. 1992. Nondestructive NIR and NIT determination of protein, fat and water in plastic wrapped, homogenized meats. Appl. Spectrosc. 46, 1685.

Isaksson, T. and Naes, T. 1988. The effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. Appl. Spectrosc. 42(7), 1273-1284.

Jasansky, A and Bilanski, W.K. 1973, Spectral absorbance of soybeans and its importance for efficient heat processing. Can. Inst. Food Sci. Technol. J. 6(3), 151-155.

Jenke, D.R. 1997. Chromatographic method validation; A review of current practices and procedures I. General concepts and guidelines. Instrum. Sci. Technol. 25, 345-359.

Jones, D.B. 1931. Factors for converting percentages of nitrogen in foods and feeds into percentages of nitrogen , USDA Circular No. 183, Washington DC.

Joslyn, M.A. 1970. Methods in Food Analysis, 2nd ed. Academic Press, New York.

Kennedy, I.R., Mwandemele, O.D. and McWhirter, K.S. 1985. Estimation of sucrose, raffinose and stachyose in soybean (Glycine max) seeds. Food Chem. 17, 85-94.

Kilen, T.C. and Kou, T.M. 1994. Soybean quantitative trait loci and marker assisted breeding. Agron. Abstr. p. 107.

Klempir J.M. 1999. Nuclear Magnetic Resonance Characterization of Soybeans For Oil Content Measurement. MS Thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois.

Knudsen, I.M. 1986. High Performance Liquid Chromatography Determination of Oligosaccharides in Leguminous Seeds. J. Sci. Food Agric. 37, 560-566.

Koga, T., Inoue, N., Asai, T. and Simizu, S. 1992. Selection of wavelength in near IR spectrometry affects accuracy of tofu lee components estimation. Chikusan no Kenkyu (in Japanese) 46(7), 773-778.

Kumosinski, T.F., Unruh, J.J. 1994. Global-secondary-structure analysis of proteins in solution. Resolution-enhanced deconvolution Fourier transform infrared spectroscopy in water. ACS Symposium Series, Philadelphia, PA. 576 (Molecular Modeling), 71.

Kyncl, J. 1982. The role of cereal production and processing Czechoslovakia, Progress in Cereal Chemistry and Technology, 1-8, Elsevier 1983.

Lamb, D. and Hurburgh, C.R. 1991. Moisture determination in single soybean seeds by near-infrared transmittance. Trans. ASAE. 34, 2123-2129.

Lanser, A.C., List, G.R., Holloway, R.K. and Mounts, T.L. 1991. FTIR estimation of free fatty acid content in crude oils extracted from damaged soybeans. J. Am. Oil Chem. Soc. 68(6), 448.

Lee, S.J., Ahn, J.K., Kim, S.H., Kim, J.T., Han, S.J., Jung, M.Y. and Chung, I.M. 2003. Variation in isoflavone of soybean cultivars with location and storage duration. J. Agric. Food Chem. 51(11), 3382-3389.

Liu, K. 1997. Soybeans Chemistry, Technology and Utilization. Chapman and Hall, Chapters 1 and 2.

Liu, K.S., Orthoefer, F. and Thompson, K. 1995. The case for food-grade soybean varieties. INFORM 6(5), 593-599.

Lowry, O.H., Rosebrough, N.J., Farr, A.L. and Randall, R.J. 1951. Protein measurement with the Folin phenol reagent. J. Bio. Chem. 193, 265-275.

Madden, H.M., 1978. Comments on the Savitzky-Golay Convolution Method for Least-Squares Fit Smoothing and Differentiation of Digital Data. Anal. Chem. 50, 1383.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. Proc. Natl. Inst. Of Science of India, 2, 49.

Mansur, L.M., Orf, J.H., Chase, K., Jarvik, T., Cregan, P.B. and Lark, K.G. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36, 1327- 1336.

Mark, H. 2001. Near-Infrared Applications in Biotechnology, Chapter 11 "Fundamentals of Near Infrared Spectroscopy", p.312, Editor Raghavachari, R. Promega Corporation, Madison, Wisconsin.

Martens, H., Jensen, S.A. and Geladi, P. 1983. Proceedings, Nordic Symposium on Applied Statistics, Stavanger, June 1983, pp.205-234.

Martin, S.K., Dye, B.W. and McBlain, B.A. 1990. Use of Hill and Short-Row Plots for Selection of Soybean Genotypes. Crop Sci. 30 (1), 74-79.
Mitchell, J. Jr., and D. M. Smith. 1948. Aquametry. John Wiley & Sons, New York.

Mora-Gutierrez, A. and Baianu, I.C. 1991. Comparisons with potato starch and potato maltodextrins. J. Agric. Food Chem. 39, 1057-1062.

Morr, C.V. 1988. Soybean utilization alternatives: a symposium sponsored by the Center for Alternative Crops and Products. pp 107-115.

Mujoo, R., Trinh, D.T. and Ng, K.W. 2003. Characterization of storage proteins in different soybean varieties and their relationship to tofu yield and texture. Food Chem. 82(2), 265-273.

Myoung, G.C., In, Y.B., Sung, T.K., Won, Y.H., Doo, C.S., Huhn, P.M. and Kwang, H.K. 2001A. Determination of protein and oil contents in soybean seed by near infrared reflectance spectroscopy. Korean J Crop Sci 46(2), 106-111.

Myoung, G.C., In, Y.B., Sung, T.K., Won, Y.H., Doo, C.S., Huhn, P.M. and Kwang, H.K. 2001B. Non-destructive method for selection of soybean lines contained high protein and oil by near infrared reflectance spectroscopy. Korean J Crop Sci 46(5), 401-406.

Nakamichi, K, Suehara, K.I., Nakano, Y., Kakugawa, K., Tamai, M., and Yano, T. 2002. Measurement of the concentrations of mannosyl erythritol lipid and soybean oil in the glycolipid fermentation process using near infrared spectroscopy. Japan. J. Near Infrared Spectrosc. 10(1), 53-61.

Nieh C.D. and Snyder, H.E. 1991. Solvent extraction of oil from soybean flour: I. Extraction rate, a countercurrent extraction system, and oil quality. J. Am. Oil Chem. Soc. 68(4), 246-249.

Nielsen, S.S. 1994. Introduction to the Chemical Analysis of Foods. Jones and Bartlet Publishers, Boston, London, Chapters 7, 10, 14.

Norris, K.H. and Williams, P.C. 1984. Optimization of mathematical treatment of raw near-IR signal in the measurement of protein in hard red spring wheat: 1. Influence of particle size. Cereal Chem. 62, 158-165.

Norris, K.H. 1983. Using gap derivatives as pre-processing for quantitative models, Food Research & Data Analysis, Proceedings of the 1982 IUFST Symposium, Martens, H. Ed., Applied Science Publishers, Oslo, 46-47.

Nzai, J. M. and Proctor, A. 1998. Determination of phospholipids in vegetable oil by Fourier transform infrared spectroscopy. J. Am. Oil Chem. Soc. 75(10), 1281-1289.

Oh, E.K., and Grossklaus, D. 1995. Measurement of the components in meat patties by near infrared reflectance spectroscopy. Meat Sci. 41, 157.

Openshaw, S.J. and Hadley, H.H. 1978. Maternal Effects on Sugar Content in Soybean Seeds. Crop Sci. 18, 581-584.

Openshaw, S.J. and Hadley, H.H. 1981. Selection to Modify Sugar Content of Soybean Seeds. Crop Sci. 21, 805-808.

Openshaw, S.J. and Hadley, H.H. 1984. Selection Indexes to Modify Protein Concentration of Soybean Seeds. Crop Sci. 24, 1-4.

Orman, B.A. and Schumann, R.A. 1992. Nondestructive single-kernel oil determination of maize by near-infrared transmission spectroscopy. J.Am.Oil Chem. Soc. 69(10), 1036-1038.

Osborne, B. G., Fearn, T. and Hindle, P.H. 1993. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. pp 13-21, 99-117.

Ozaki Y., Sasic, S. and Jiang, J. 2001. How can we unravel complicated near infrared spectra? – recent progress in spectral analysis methods for resolution enhancement and band assignment in the near infrared region. J. Near Infrared Spectrosc. 9, 63-95.

Parreira, T.F., Ferreira, M.C., Sales, H.J. and DeAlmeida, W.B. 2002. Quantitative determination of epoxidized soybean oil using near-infrared spectroscopy and multivariate calibration. Appl. Spectrosc. 56(12), 1607-1614.

Pazdernik, D.L., Killam, A.S. and Orf, J.H. 1997. Analysis of amino and fatty acid composition in soybean seed, using near infrared reflectance spectroscopy. Agron. J. 89(4), 679-685.

Pazdernik, D.L., Plehn, S.J., Halgerson, J.L. and Orf, J.H. 1996. Effect of Temperature and Genotype on the Crude Glycinin Fraction (11S) of Soybean and Its Analysis by Near-Infrared Reflectance Spectroscopy (Near-IRS). J. Agric. Food Chem. 44 (8), 2278 –2281.

Peris-Tortajada, M. 2000. HPLC Determination of Carbohydrates in Foods, Chapter 7, Food Analysis by HPLC, 2nd ed. Edited by Leo M. L. Nollet, Marcel Dekker Inc., New York

PerkinElmer Instruments. 2000. Spectrum One User's Guide, Shelton, CT 06484, USA.

Peterson, G.L. 1979. Review of the Folin phenol protein quantitation method of Lowry, Rosebrough, Farr and Randall. Analytical Biochemistry. 100, 201-220.

Pierce, M.M. and Wehling R.L. 1994. Comparison of sample handling and data treatment methods for determining moisture and fat in Cheddar cheese by near infrared spectroscopy. J. Agric. Food Chem. 42, 2830.

Pfeffer, P. and Gerasimowicz, W. (Eds.). 1987. NMR in Agriculture. CRS Publishers: Boca Raton, FL.

Pomeranz Y. and Meloan, C.E. 1994. Food Analysis Theory and Practice, third edition, chapter 33, 567-574. Chapman & Hall, New York, NY.

Psotka, J. and Shadow, W. 1994. NIR analysis in the wet corn refining industry – A technology review of methods in use. Intern. Sugar J. 96, 358.

Rasco, B.A., Miller, C.E. and King, T.L. 1991. Utilization of NIR spectroscopy to estimate the proximate composition of trout muscle with minimal sample pretreatment. J. Agric. Food Chem. 39, 67.

Reffer, J.A. and Martoglio, P.A. 1995. Uniting Microscopy and Spectroscopy, Practical Guide to Infrared Microspectroscopy, Chapter 2, p.41, Edited by Howard J. Humecki, Marcel Dekker, Inc.

Rinne, R.W., Gibbons, S., Bradley, J., Seif R. and Grim, C.A. 1975. Soybean Protein and Oil percentages Determined by Infrared Analysis. USDA-ARS-NC-26, 1-4.

Robinson, H.W. and Hodgen, C.G. 1940. The biuret reaction in the determination of serum protein. 1. A study of the conditions necessary for the production of the stable color which bears a quantitative relationship to the protein concentration. J. Bio. Chem. 135, 707-725.

Rodriquez-Otero, J.L., Hermida, M. and Cepeda, A. 1995. Determination of fat, protein and total solids in cheese by near infrared reflectance spectroscopy. J. AOAC International. 78, 802.

Rodriguez-Saona, L. E., Fry F.S. and Calvey, E.M. 2000. Use of Fourier Transform Near-Infrared Reflectance Spectroscopy for Rapid Quantification of Castor Bean Meal in a Selection of Flour-Based Products. J. Agric. Food Chem. 48(11), 5169-5177.

Sato, T, Abe, H., Kawano, A., Ueno, G., Suzuki, K. and Iwamoto, M. 1994. Near-Infrared Spectroscopic Analysis of Deterioration Indices of Soybeans for Process Control in Oil Milling Plant. J. Am. Oil Chem Soc. 71(10), 1049-1055.

Savitsky, A. and Golay, M.J. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627-1639.

Sehern, N.A. and Lambert, J.W. 1984. Effect of stratification for percent protein in two soybean populations. Crop Sci. 24, 225-228

Shadow, W. 1998. Rapid Analysis for the Food Industry Using Near-Infrared Spectroscopy. Perten Instruments North America, Inc., 1-17.

Silvela, L., Rodgers, R., Barrera, A. and Alexander, D.E. 1989. Effect of selection intensity and population size on percent oil in maize, Zea mays L. Ther. Appl. Genet. 78(2), 298-304.

Simpson, A.M. Jr. and Wilcox, J.R. 1983. Genetic and phenotypic associations of agronomic characteristics in four high protein soybean populations [Glycine max]. Crop Sci. 23, 1077.

Smith, K.J. and Huyser, W. 1987. In Soybeans: Improvement, Production, and Uses (J.R. Wilcox ed.), p.19.

Society for The Study of Natto. 1990. Methods of natto research. 1st ed. (In Japanese). National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries, Japan.

Song, T., Barua, K., Buseman, G. and Murphy, P.A. 1998. Soy isoflavone analysis: quality control and a new internal standard. Am. J. Clinical Nutrition. 68 (6), supplement, 1474S-1479S.

Specht, J., Nelson, R.L., Webster, R., Naidu, S., Ainsworth, E. and Ort, D. 2002. A genomic perspective on the soybean protein(+)/oil(-)/yield(-) enigma. Program and Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean, August 11-14, 2002, P.202.

Steiner, J., Termonia, Y. and Deltour, J. 1972. Comments on smoothing and di€erentiation of data by simplified least square procedure. Anal. Chem. 44, 1906-1909.

Stuber, C.W. 1992, Biochemical and Molecular Markers in Plant Breeding. Plant Breeding Rev. 9, 37-61.

Tajuddin, T., Watanabe, S., Masuda, R., Harada, K. and Kawano, S. 2002. Application of near infrared transmittance spectroscopy to the estimation of protein and lipid contents in single seeds of soybean recombinant inbred lines for quantitative trait loci analysis. J. Near Infrared Spectrosc. 10(4), 315-325.

Tekel, J., Daeseleire, E., Heeremans, A. and Peteghem, C.V. 1999. Development of a simple method for the determination of genistein, daidzein, biochanin A, and formononetin (biochanin B) in human urine. J. Agric. Food Chem. 47(9), 3489-3494.

Thanh, V.H. and Shibasaki, K.J. 1976. Major proteins of soybean seeds. A straight-forward fractionation and their characterization. Agric. Food Chem. 24(6), 1117-1121.

Tomomatsu, H. 1994. Health effects of oligosaccharides. Food Technol. Oct, 61-65.

Wang, D., Dowell, F.E. and Lacey, R.E. 1999. Single wheat kernel color classification by using near-infrared reflectance spectra. Cereal Chem. 76(1), 30-33.

Wang, X., Zhao, S. and Wang, R. 2001. Analysis of soybean lecithin by supercritical fluid chromatography. Sepu. 19(4), 344-346 (Journal written in Chinese).

Wehling, R.L., Pierce, M.M. and Froning, G.W. 1988. Determination of moisture, fat and protein in spray dried whole egg by near infrared reflectance spectroscopy. J. Food Sci. 53, 1356.

Wesley, I.J., Osborne, B.G., Anderssen, R.S and Skerritt, J.H. 1999. Curve fitting applied to near infrared deconvolution of wheat functional proteins in flour. Near Infrared Spectroscopy, Proceedings of the International Conference, 9th, Verona, Italy, p.215.

Wehrmann, V.K., Fehr, W.R., Cianzio, S.R. and Cavins, J.F. 1987. Transfer of high seed protein to high yielding soybean cultivars. Crop Sci. 25, 927-931.

Whistler, R.L. and BeMiller, J.N. 1997. Carbohydrates Chemistry for Food Scientists, Eagen Press, St. Paul, MN.

Wilcox, J.R. 1998. Increasing seed protein in soybean with eight cycles of recurrent selection. Crop Sci. 38, 1536-1540.

Wilcox, J.R. and Cavins, J.F. 1995. Backcrossing high seed protein to a soybean cultivar. Crop Sci. 35, 1036-1041.

Wilcox J.R. and Shibles, R.M. 2001. Interrelationship among Seed Quality Attributes in Soybean, Crop Sci. 41, 11-14.

Wilson, J.M., Kramer, A. and Ben-Gera, I. 1973. Quantitative Determination of Fat, Protein and Carbohydrates of Soy Products with Infrared Attenuated Total Reflectance. J. Food Sci. 38(1), 14-17.

Williams, P. and Norris, K. 1987. Near-Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, Inc. St.Paul, MN. pp 43-47.

Williams, P.C. and Sobering, D.C. 1993. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. J. Near Infrared Spectrosc. 1, 25-32.

Wilson, J.M., Kramer, A. and Ben-Gera, I. 1973. Quantitative determination of fat, protein, and carbohydrates of soy products with infrared attenuated total reflectance. J. Food Sci. 38(1), 14-17.

Wilson, R.F. 1991. Designing Value-Added Soybeans for Markets of the Future. *American Oil Chemists' Society Publication*, Champaign, Illinois, pages 17, 53, 80, 88, 109, 115.

Xu, H. and Wilcox, J.R. 1992. Recurrent selection for maturity and percent seed protein in Glycine max based on S0 plant evaluations. *Euphytica 62*, 51-57.

Yazdi-Samadi B., Rinne, R.W., and Seif, R.D. 1977. Components of Developing Soybean Seeds: Oil, Protein, Sugars, Starch, Organic Acids and Amino Acids. *Agron. J*. 69, 481-486.

You, T., Guo, J., Baianu, I.C. and Nelson, R.L. 2002A. Determination of Isoflavones Contents for Selected Soybean Lines by Fourier Transform Near Infrared Reflectance Spectroscopy". *Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean*, August 11-14, 2002, P505.

You, T., Guo, J., Baianu, I.C. and Nelson, R.L. 2002B. Calibration of dual-diode array and fourier transform near infrared reflectance spectrometers for composition analysis of single soybean seeds in genetic selection, cross-breeding experiments. *Program and Proceedings for the 9th Biennial Conference of the Cellular and Molecular Biology of the Soybean*, August 11-14, 2002, page 508.

You T., Guo, J., Baianu, I.C., Nelson, R.L. 2002C. Rapid Determination of Protein, Oil, Moisture, and Isoflavones Contents of Single Soybean Seeds by Fourier Transform Near Infrared Reflectance Spectroscopy. Proceedings of the China and International Soy Conference and Exhibition 2002 (CISCE 2002), November 6-9, 2002, 414-415.

You, T., Baianu, I.C., Guo, J., Nelson, R.L. 2011a. Diode-Array Near Infrared Spectroscopy for Rapid Analysis of Soybeans: Light Scattering Corrections for Intact and Ground Soybean Seeds. (*submitted*).

You T., Guo, J., Baianu, I.C. and Nelson, R.L. 2011b, The impact of soybean composition ranges on calibration development with PLS-1 algorithm by theoretical simulations, . (*submitted*)

CHAPTER 2